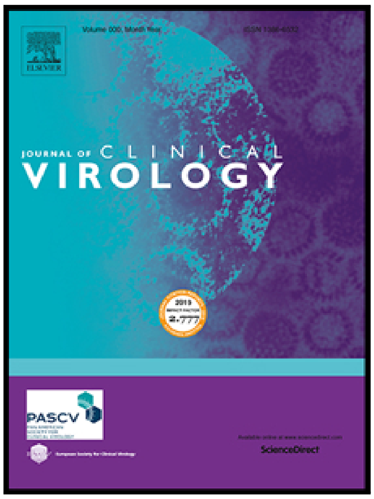
Journal Pre-proof



A metagenomics workflow for SARS-CoV-2 identification, co-pathogen detection, and overall diversity

Daniel Castañeda-Mogollón , Claire Kamaliddin , Lisa Oberding , Yan Liu , Abu Naser Mohon , Rehan Mujeeb Faridi , Faisal Khan , Dylan Pillai

PII: S1386-6532(21)00292-4
DOI: https://doi.org/10.1016/j.jcv.2021.105025
Reference: JCV 105025

To appear in: *Journal of Clinical Virology*

Received date: 21 June 2021
Revised date: 28 October 2021
Accepted date: 1 November 2021

Please cite this article as: Daniel Castañeda-Mogollón , Claire Kamaliddin , Lisa Oberding , Yan Liu , Abu Naser Mohon , Rehan Mujeeb Faridi , Faisal Khan , Dylan Pillai , A metagenomics workflow for SARS-CoV-2 identification, co-pathogen detection, and overall diversity, *Journal of Clinical Virology* (2021), doi: https://doi.org/10.1016/j.jcv.2021.105025

Highlights

- We describe the first study to apply a clinical metagenomic pipeline to identify SARS-CoV-2 variants of concern.
- The SARS-CoV-2 variants of concern (B.1.1.7) and 2 variants of interest (P.2) were successfully identified.
- Several bacterial co-pathogens were noted in the SARS-CoV-2 infected patients.
- The bioinformatic and experimenal pipeline developed here presents an important advancement in unbiased diagnostic approaches to identify and define pandemic viruses.

Title

A metagenomics workflow for SARS-CoV-2 identification, co-pathogen detection, and overall diversity

Authors

Daniel Castañeda-Mogollón[1,2,3], Claire Kamaliddin[1,2,3], Lisa Oberding[1,2,3], Yan Liu[1,2,3], Abu Naser Mohon[1,2], Rehan Mujeeb Faridi[4,5,6], Faisal Khan[4,5,6], Dylan Pillai[1,2,3,4]

Affiliations

1. Cumming School of Medicine, Department of Pathology & Laboratory Medicine, the University of Calgary, AB, Canada

2. Cumming School of Medicine, Department of Microbiology, Immunology, and Infectious Diseases, the University of Calgary, Canada.

3. Calvin, Phoebe & Joan Snyder Institute for Chronic Diseases, the University of Calgary, Calgary, AB, Canada

4. Alberta Precision Laboratories, Diagnostic & Scientific Centre, Calgary, AB, Canada

5. Hematology Translational Lab, University of Calgary, Calgary, AB, Canada

6. Arnie Charbonneau Cancer Institute, the University of Calgary, Calgary, AB, Canada

Corresponding author

Dylan R. Pillai MD, PhD, FRCP(C)

Departments of Pathology & Laboratory Medicine, Medicine, and Microbiology & Infectious Diseases,

& Community Health Sciences University of Calgary,

Diagnostic & Scientific Centre,

Room 1W-416,

9-3535 Research Road N.W.

Calgary, AB T2L 2K8

Ph: 403-770-3578

drpillai@ucalgary.ca

Abstract

An unbiased metagenomics approach to virus identification can be essential in the initial phase of a pandemic. Better molecular surveillance strategies are needed for the detection of SARS-CoV-2 variants of concern and potential co-pathogens triggering respiratory symptoms. Here, a metagenomics workflow was developed to identify the metagenome diversity by SARS-CoV-2 diagnosis ($n_{positive}$=65; $n_{negative}$= 60), symptomatology status ($n_{symptomatic}$=71; $n_{asymptomatic}$=54) and anatomical swabbing site ($n_{nasopharyngeal}$=96; n_{throat}=29) in 125 individuals. Furthermore, the workflow was able to identify putative respiratory co-pathogens, and the SARS-CoV-2 lineage across 29 samples. The diversity analysis showed a significant shift in the DNA-metagenome by symptomatology status and anatomical swabbing site. Additionally, metagenomic diversity differed between SARS-CoV-2 infected and uninfected asymptomatic individuals. While 31 co-pathogens were identified in SARS-CoV-2 infected patients, no significant increase in pathogen or associated reads were noted when compared to SARS-CoV-2 negative patients. The Alpha SARS-CoV-2 VOC and 2 variants of interest (Zeta) were successfully identified for the first time using a clinical metagenomics approach. The metagenomics pipeline showed a sensitivity of 86% and a specificity of 72% for the detection of SARS-CoV-2. Clinical metagenomics can be employed to identify SARS-CoV-2 variants and respiratory co-pathogens potentially contributing to COVID-19 symptoms. The overall diversity analysis suggests a complex set of microorganisms with different genomic abundance profiles in SARS-CoV-2 infected patients compared to healthy controls. More studies are needed to correlate severity of COVID-19 disease in relation to potential disbyosis in

the upper respiratory tract. A metagenomics approach is particularly useful when novel pandemic pathogens emerge.

Keywords: Metagenomics, SARS-CoV-2, COVID-19, variants of concern, variants of interest.

Introduction

Individuals infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) present with multiple symptoms ranging in severity from asymptomatic/mild cases to severe pneumonia and death[1]. The poor specificity of COVID-19 clinical presentation means that extensive screening must be performed for individuals presenting fever or respiratory infection symptoms. Current screening strategies are based on nasopharyngeal swabs (NPS) or throat swabs (TS) and molecular diagnostics targeting specific SARS-CoV-2 genes. Recently, multiple SARS-CoV-2 variants have been identified through whole-genome sequencing (WGS) approaches, including variant of concern (VOC) Alpha, Beta, Gamma, Epsilon, and Delta[2].

Metagenomic Next-Generation Sequencing (mNGS) provides an unbiased method for identification of all taxonomic ranks in a sample using a single sequencing run[3,4]. Compared to traditional microbial culture-based methods, mNGS can be used as a robust diagnostic tool, which is faster, more sensitive, and allows for the identification of unculturable organisms[5–7]. At the time of writing, the metagenome associated with SARS-CoV-2 infection remains poorly characterized. Additionally, current COVID-19 co-infection studies could be biased by public health guidelines (i.e. social distancing, masks)[8–14,] as pre-pandemic studies have found *S. pneumoniae*, *P. aeruginosa*, and *H. influenzae* to be the three most common bacterial co-pathogens[15]. While the aforementioned studies offer clues to understand the microbial diversity associated with COVID-19, a study of the metagenome and metatranscriptome (hereinafter referred as DNA-metagenome and RNA-metagenome) associated with COVID-19 and SARS-CoV-2 asymptomatic infection is necessary. This study investigated the metagenome from upper respiratory samples by SARS-Cov-2 diagnosis, symptomatology, and anatomical sampling site (Figure 1). Moreover, we evaluated the performance of mNGS for SARS-CoV-2 diagnosis, and its ability to identify SARS-CoV-2 mutants.

Material and methods

2.1 Sample collection

A total of 125 clinical NPS and TS samples were collected and tested by Alberta Precision Laboratories between March 2020 and February 2021. Swabs were performed by trained personnel as part of the Alberta COVID-19 testing program. Symptom screening was based on patient reporting to the sampling nurse using the standard APL procedure (Supplementary methods).

2.2 Ethics statement

Ethical approval was obtained from Conjoint Health Research Ethics Board (CHREB) of the University of Calgary (REB 20-0567, REB 20-0402). All archived specimens were de-identified prior to analysis in this study. Informed consent was waived by the ethics board.

2.3 Nucleic acid extraction

Samples were randomized in extraction batches including internal controls to assess the metagenome kitome. DNA and RNA were extracted using the Qiagen QIAamp® DNA Mini Kit (Cat. No./ID: 51306, Qiagen, Germany) and the Qiagen QIAamp® Viral RNA Mini Kit (Cat. No/ID 52906, Qiagen, USA) respectively. Both protocols were adapted from the manufacturer's recommendation (Supplementary Methods).

2.4 cDNA synthesis

Primer spiked enrichment was adapted from published protocols[16,17]. cDNA synthesis was performed from 5 µL of extracted RNA (DNA-free) using the NEBNext Ultra II first strand and second strand synthesis modules (E7771 and E6111, NEB, MA, USA) (Supplementary Methods).

2.5 Internal controls, library preparation and sequencing

Internal controls were used to assess the overall performance of the mNGS pipeline, as well as to generate a background model to remove environmental contamination. The product from the cDNA synthesis step were used in the library preparation step, and sequenced in a Illumina instrument (Illumina, USA) using a NovaSeq 300 cycle SP v1.5 kit set (Illumina, USA) for 2 x 150 bp paired-end (detailed in Supplementary Methods).

2.6 Metagenome description and identification of infectious agents

Organism detection was performed using sequence analysis of metagenomic data using the IDseq server-based pipeline [18]. The quality control step performed *a priori* subtraction of host sequences by using STAR (Spliced Transcripts Alignment to a Reference) [19], followed by Trimmomatic [20] to trim Illumina adapters. Low-quality and low-complexity reads were removed followed by taxonomic identification (detailed in Supplementary Methods). Two filters were applied to increase the analytical specificity of the workflow for species identification: (i) a Z-score $\geq$ 2.0, and (ii) a minimum of 10 reads aligned to the NCBI Nucleotide database. To determine sample diversity, an alpha- and beta-diversity analysis was performed across each group (detailed in Supplementary Methods). In addition a diversity analysis was stratified by SARS-CoV-2-positive samples with Ct values above 30.

2.6 SARS-CoV-2 genome assembly and variant calling

Samples with reads mapped to the SARS-CoV-2 genome were submitted to IDseq for genome assembly and variant calling. SNPs were called for variation analysis and compared against the reference genome using the default parameters. For sample lineage characterization, genomes with a minimum breadth of coverage of 50% were submitted to the Pangolin online sequence aligner[21] (based on the GISAID consortium https://www.gisaid.org/ - available sequences on March 27th, 2021). Lineage characterization for samples between 25% to 50% breadth of coverage were estimated by the closest clade in the phylogenomics tree (Supplementary Methods).

2.8 Identification of putative respiratory pathogens

Species were identified based on Z-score ≥ 2 and > 10 reads mapped to a given taxa. The complete list of species that were screened as part of the putative respiratory pathogen panel is available in Supplementary Methods. Proportion of identified organism were compared using Fisher's exact test with Benjamini-Hocheberg correction.

2.9 Statistical analysis

Details of the statistical analysis and software are available in the Supplementary Methods.

Results

3.1 Patient population

A total of 125 samples (96 nasopharyngeal swabs [NPS], and 29 throat swabs [TS]) were included in the study. Seventy one patients were symptomatic and 54 asymptomatic. A total of 65/125 samples were positive for SARS-CoV-2 by E-gene RT-PCR performed by the clinical laboratory[22] .

3.2 Assessment of the respiratory metagenome

A total of 823,317,205 and 765,758,597 non-human reads were sequenced from the metagenome and metatranscriptome, respectively (1.07 DNA to cDNA ratio). An average of 20,983,714 ± 358,651 and 20,479,932 ± 469,732 reads were identified respectively for the DNA-metagenome and the RNA-metagenome. No significance was found amongst the non-human reads by SARS-CoV-2 diagnosis (Figure S1a) nor by anatomical sampling site (Figure S1b). Significantly higher human reads were observed amongst NPS than TS in the cDNA number of reads (Figure S1c). Amongst the SARS-CoV-2 infected individuals, an average of 0.02% of reads were mapped to the SARS-CoV-2 genome from the original number of raw cDNA reads.

The DNA-metagenome diversity analysis of significance was performed (Table 1).The DNA-metagenome beta-diversity (Figure 2a-f) showed significant results in the quantitative (Bray-Curtis metric) and qualitative (Jaccard metric) analysis by symptomatology status, and anatomical swabbing site. Significant results were also observed by the quantitative beta-diversity PCoA plot amongst the asymptomatic NPS samples by SARS-CoV-2 diagnosis status (Figure 2d); The NPS-asymptomatic sub-cohort by SARS-CoV-2 showed significant results by its qualitative analysis (Figure 2e). The alpha-diversity analysis in the Shannon index showed significance by the Wilcoxon-ranked test by anatomical swabbing site and symptomatology status but not by the remaining analysis (Figure 2 g-l). The DNA-metagenome diversity analysis by SARS-CoV-2 diagnosis after excluding SARS-CoV-2positive samples with Ct values above 30 did not show significant results in the alpha- and beta-diversity analysis (Table 1).

The diversity analysis of significance was also performed for the RNA-metagenome of bacteriophages (Table S1). The beta-diversity analysis of the bacteriophages RNA-metagenome showed significant results in its quantitative and qualitative analysis by anatomical sampling site, symptomatology, and the SARS-CoV-2 status amongst the NPS-asymptomatic cohort (Figure 3a-f). The alpha-diversity analysis in the Shannon index showed significance by SARS-CoV-2 diagnosis status, symptomatology, and anatomical swabbing site (Figure 3g-l). The RNA-metagenome for RNA-viruses showed 4 microorganisms, including *Enterovirus D, Influenza A, Rhinovirus,* and uncultured virus. No significant differences by RNA-viruses in terms of abundance or presence/absence were observed by anatomical swabbing site, symptomatology status, or by SARS-CoV-2 diagnosis. A relative abundance analysis across the NPS samples identified 203 species from various domains (bacteria, archaea, eukarya, DNA-viruses, DNA-bacteriophages; table S2) with a significant fold change between COVID-19 and healthy NPS (Figure 4a). Amongst these species, only one DNA-virus was identified (*Human betaherpes virus 6*; Supplementary file 1). No significant species were detected between NPS-asymptomatic SARS-CoV-2 positive vs NPS-asymptomatic SARS-CoV-2 negative (Figure 4b). No significant species were found amongst NPS-symptomtic by SARS-CoV-2 diagnosis status (Figure 4c).

3.3 Identification of putative respiratory pathogens

The DNA and RNA-metagenome results were screened for presence of potential pathogens. A total of 31 pathogens were identified across the samples from the respiratory pathogen panel. Seventeen (17/31) pathogens were identified in at least one sample. Fourteen and nine organisms of interest were detected in the NPS and TS samples, respectively . In the TS, three microorganisms were unique to SARS-CoV-2 negative patients (*Streptococcus pyogenes, Serratia marscescens,* and *Dolosigranolum pigrum),* and the remaining seven identified were found in both positive and negative patients. Among the NPS, five unique microorganisms were detected in SARS-CoV-2 positive patients (*Moraxella catharralis, Klebsiella pneumoniae*, human betaherpes virus 6, *Haemophilus parainfluenzae*, and *Dolosigranulum prigrum*) and seven were unique to negative patients *(*Rhinovirus, *M. pneumoniae,* Influenza A virus, *H.*

influenzae, C. pneumoniae, human coronavirus HKU1 and NL63). The prevalence of each screened potential co-pathogen was compared between SARS-CoV-2 positive (COVID-19 patients) and uninfected patients for NPS (Table S3) and TS (Table S4). Only *D. pigrum* was significantly more prevalent in COVID-19 positive patients. No significant rPM differences were observed for the rest of the microorganisms in NPS or TS (Figure 5).

3.4 SARS-CoV-2 detection by mNGS

A negative relationship was observed between aligned viral reads and corresponding RT-PCR E-gene Ct value (R^2=0.45) using an exponential regression model (Figure S2a). Significant correlation (p<0.0001) between the SARS-CoV-2 mapped reads and the E-gene Ct value was observed (Spearman's ρ =-0.77, Pearson's r=-0.53,). The interpolation of the exponential regression model suggests a Ct-value of 37.19 for the detection of 25 reads, and a Ct-value of 39.05 for 5 SARS-CoV-2 reads. Similar exponential models were generated by NPS and TS (Figure 6a). NPS had a higher rate of SARS-CoV-2 DNA read retrieval than TS. A logistic regression model was generated for genome coverage and number of SARS-CoV-2 mapped reads, with an excellent fit (R^2=0.98) (Figure S2c). The presented model suggests an approximate of 2,500 and 7,900 reads needed to assemble 50% and over 98% of the SARS-CoV-2 genome, respectively.

Receiver operating characteristic (ROC) curves of SARS-CoV-2 mapped reads (Figure 6b) showed an analytical sensitivity of 0.71 (95%CI=[0.58,0.82]), a specificity of 0.86 (95%CI=[0.72,0.93]), and an area under the curve (AUC) of 0.85 (95%CI=[0.77,0.92]) for NPS. An analytical sensitivity of 0.91

(95%CI=[0.64,0.99]), and a specificity of 0.70 (95%CI=[0.46,0.86]) for TS. A total of 25 SARS-CoV-2 reads was calculated to be the optimum number of reads to achieve the highest sensitivity and specificity of any clinical sample, regardless of the anatomical sampling site (Figure S2b).

3.5 SARS-CoV-2 genomic variation and lineage identification

A total of 274 single nucleotide polymorphisms (SNPs) and deletions were identified, of which 128 were unique. The majority (63.28%, n=81/128) corresponded to non-synonymous mutations. The remaining SNPs were either synonymous (26.56%, n=34/128), deletions (3.90%, n=5/128), nonsense (2.34%, n=3/128), or located in the non-coding regions (3.90%, n=5/128). The majority were observed in the ORF1ab gene (Figure 7b). The mNGS pipeline identified 26 SNPs that are signatures of VOCs/VOIs in three samples (P739, P743, and P744) (Figure 7a). Two samples (P743 and P744) contained the Zeta VOI, and one sample (P739) contained the Alpha VOC. The Alpha isolate presented 16/17 SNPs and deletions that characterize this VOC [(https://cov-lineages.org/global_report_B.1.1.7.html)]. The Zeta positive samples displayed five and 11 characteristic SNPs out of the 13 lineage-defining mutations, respectively. Among the 128 unique SNPs and deletions identified by mNGS, 28.90% (n=37/128) have annotated features and/or predicted changes that differ from the wild-type virus (Table S5). A total of 36 out of 65 SARS-CoV-2 positive samples (55.38%) had a WGS and S gene coverage below 50%.

Twenty-nine SARS-CoV-2 genomes were properly identified.The majority of the samples were assigned to the B lineage (24/29). Two samples were classified as part of the Zeta VOI (2/29) with a breadth of coverage of 98.2% and 99.9%, respectively (Figure S3a, S3b). One sample was classified as Alpha VOC (1/29) with a breadth of coverage of 98.3% (Figure S3c). One sample was assigned as part of the D lineage (1/29), and one as the A.1 lineage (Figure 8). Out of this, 23 samples were properly identified using the phylogenomics tree generated and the PANGO-Lineage assigner. Overall, the mNGS workflow identified one VOC and two samples with a single VOI (Table S6).

Discussion

This study has provided evidence that the mNGS workflow can detect a significant shift in the overall metagenome variability. The metagenome PCoA diversity analysis revealed no significant metagenome variability by SARS-CoV-2 infected status in the overall DNA-metagenome, however, significant findings were found amongst the RNA bacteriophage by SARS-CoV-2 diagnosis status; these results are discordant with previous quantitative-PCA reports on the NPS bacterial microbiome[8]. Similarly, Han *et al.* reported significant quantitative bacteriome and virome differences by PCoA between the bronchoalveolar lavage fluid (BALF) of SARS-CoV-2 infected and non-infected patients[23]. These results are in partial agreement with the findings of Rosas-Salazar *et al*[14] .Similarly, our results agree with the Shannon index obtained by previous studies[14,24].

Importantly, mNGS also allows for the unbiased identification of co-pathogens or other infectious aetiologies in samples. In the SARS-CoV-2 negative symptomatic patients, these pathogens are clinically relevant for upper respiratory infection symptoms (rhinovirus, *M. pneumoniae, C.*

pneumoniae, influenza or other coronaviruses). Screening and detection of other pathogens may be in favour of co-infections among the COVID-19 positive patients, as reported elsewhere[9–12]. Previous meta-analysis reported higher proportion of bacterial co-infection in the intensive care unit (ICU) patients, reflecting disease severity [25], but these co-infections may also be related to the level of care[26]. The presented study did not confirm the presence of *M. pneunomiae*, *P. aeruginosa* and *H. influenzae* as identified elsewhere[25]. In addition, the mNGS workflow did not identify any fungal co-infection among the screened organism, as reported elsewhere based on clinical laboratory findings[27].

In terms of lineage, the majority (n=26/29) of the reconstructed SARS-CoV-2 genomes in this study were clustered in the A and B lineages. A major strength of this pipeline is VOC/VOI's identification among the studied samples, showing the potential of mNGS as a surveillance tool for VOC/VOI spread and the monitoring of new variants. Nevertheless, this workflow was able to recover the genome of 29 samples, suggesting the remaining 36 as potentially missed VOC/VOI calls. Moreover, the VOC/VOI calling is based on the SNPs of the entire SARS-CoV-2 genome. VOC/VOIs can be identified by signature SNPs in the S gene of the virus, by either using capillary sequencing[28] or amplicon deep sequencing[29]. The latter can identify variants with a higher depth in the S gene while reducing the cost of a WGS pipeline.

A weakness of the study is the low number of individual tested (n=125), nevertheless, at the time of writing, the results here reported have the highest number of analyzed samples amongst similar studies.

Overall, metagenomics sequencing can be adapted for the current ongoing COVID-19 pandemic as well as emerging viral pandemic threats.

Data availability

All sequencing results are available at the European Nucleotide Archive (ENA) (Project ID ERP132183; sequencing sample IDs ERS7669237 – ERS7669486).

Funding

This work was funded by Canadian Institutes of Health Research (NFRFR-2019-00015) and Genome Canada (Alberta, GC 2020-03-6).

Authors' contributions

D.P., D.C.M., L.O and C.K. designed the experiments, D.P. facilitated sample provision with clinical data, C.K., L.O., and A.M. performed the experiments. D.C.M. performed the downstream bioinformatics analysis. D.C.M., C.K., and Y.L. analyzed the data and prepared the figures. D.C.M., C.K., L.O., and D.P. wrote the manuscript. D.P. and L.O. reviewed the manuscript.

Acknowledgements

We would like to thank the participants of the study. We would like to thank Paul Gordon and Elizabeth Fashbender for their valuable input in the study design. We also thank technologists at Alberta Precision Laboratories for sample processing.

References

1. Esakandari, H. *et al.* A comprehensive review of COVID-19 characteristics. *Biol. Proced. Online* **22**, (2020).
2. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* (2021) doi:10.1126/science.abg3055.
3. Sawyer, A., Free, T. & Martin, J. Metagenomics: preventing future pandemics. *BioTechniques* **70**, 1–4 (2021).
4. Miller, S. *et al.* Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res.* **29**, 831–842 (2019).
5. Huang, J. *et al.* Metagenomic Next-Generation Sequencing versus Traditional Pathogen Detection in the Diagnosis of Peripheral Pulmonary Infectious Lesions. *Infect. Drug Resist.* **13**, 567–576 (2020).
6. Duan, H. *et al.* The diagnostic value of metagenomic next-generation sequencing in infectious diseases. *BMC Infect. Dis.* **21**, 62 (2021).
7. Huang, Z. *et al.* Pathogenic Detection by Metagenomic Next-Generation Sequencing in Osteoarticular Infections. *Front. Cell. Infect. Microbiol.* **10**, 471 (2020).

8. Mostafa, H. H. *et al.* Metagenomic Next-Generation Sequencing of Nasopharyngeal Specimens Collected from Confirmed and Suspect COVID-19 Patients. *mBio* **11**, (2020).
9. Kim, D., Quinn, J., Pinsky, B., Shah, N. H. & Brown, I. Rates of Co-infection Between SARS-CoV-2 and Other Respiratory Pathogens. *JAMA* **323**, 2085–2086 (2020).
10. Van Tan, L. *et al.* SARS-CoV-2 and co-infections detection in nasopharyngeal throat swabs of COVID-19 patients by metagenomics. *J. Infect.* **81**, (2020).
11. Shah, S. J. *et al.* Clinical features, diagnostics, and outcomes of patients presenting with acute respiratory illness: a comparison of patients with and without COVID-19. *MedRxiv Prepr. Serv. Health Sci.* (2020) doi:10.1101/2020.05.02.20082461.
12. Richardson, S. *et al.* Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* **323**, 2052–2059 (2020).
13. Wu, J. *et al.* Clinical Characteristics of Imported Cases of Coronavirus Disease 2019 (COVID-19) in Jiangsu Province: A Multicenter Descriptive Study. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **71**, 706–712 (2020).
14. Rosas-Salazar, C. *et al.* SARS-CoV-2 infection and viral load are associated with the upper respiratory tract microbiome. *J. Allergy Clin. Immunol.* **147**, 1226-1233.e2 (2021).
15. Liu, Y. *et al.* Outcomes of respiratory viral-bacterial co-infection in adult hospitalized patients. *EClinicalMedicine* **37**, (2021).

16. Deng, X. *et al.* Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. *Nat. Microbiol.* **5**, 443–454 (2020).

17. Manning, J. E. *et al.* Rapid metagenomic characterization of a case of imported COVID-19 in Cambodia. *bioRxiv* (2020) doi:10.1101/2020.03.02.968818.

18. Kalantar, K. L. *et al.* IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *GigaScience* **9**, (2020).

19. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

20. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30**, 2114–2120 (2014).

21. *cov-lineages/pangolin*. (CoV-lineages, 2021).

22. Pabbaraju, K. *et al.* Development and validation of RT-PCR assays for testing for SARS-CoV-2. *Off. J. Assoc. Med. Microbiol. Infect. Dis. Can.* e20200026 (2021) doi:10.3138/jammi-2020-0026.

23. Han, Y., Jia, Z., Shi, J., Wang, W. & He, K. The active lung microbiota landscape of COVID-19 patients. *medRxiv* 2020.08.20.20144014 (2020) doi:10.1101/2020.08.20.20144014.

24. De Maio, F. *et al.* Nasopharyngeal Microbiota Profiling of SARS-CoV-2 Infected Patients. *Biol. Proced. Online* **22**, 18 (2020).

25. Lansbury, L., Lim, B., Baskaran, V. & Lim, W. S. Co-infections in people with COVID-19: a systematic review and meta-analysis. *J. Infect.* **81**, 266–275 (2020).

26. Thaden, J. T. & Maskarinec, S. A. When two for the price of one isn't a bargain: estimating prevalence and microbiology of bacterial co-infections in patients with COVID-19. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* **26**, 1602–1603 (2020).

27. Hughes, S., Troise, O., Donaldson, H., Mughal, N. & Moore, L. S. P. Bacterial and fungal coinfection among hospitalized patients with COVID-19: a retrospective cohort study in a UK secondary-care setting. *Clin. Microbiol. Infect.* **26**, 1395–1399 (2020).

28. Jørgensen, T. S. *et al. A rapid, cost efficient and simple method to identify current SARS-CoV-2 variants of concern by Sanger sequencing part of the spike protein gene*. 2021.03.27.21252266 https://www.medrxiv.org/content/10.1101/2021.03.27.21252266v1 (2021) doi:10.1101/2021.03.27.21252266.

29. Fass, E. *et al. HiSpike: A high-throughput cost effective sequencing method for the SARS-CoV-2 spike gene*. 2021.03.02.21252290 https://www.medrxiv.org/content/10.1101/2021.03.02.21252290v1 (2021) doi:10.1101/2021.03.02.21252290.

Tables

Table 1. Statistical analysis of the alpha and beta diversity results of the DNA-metagenome. The alpha-diversity analysis was evaluated using a pairwise Kruskal-Wallis test. The beta-diversity analysis was

assessed using a permutational multivariate analysis of variance (PERMANOVA). Results with a p-value < 0.05 (bolded) were considered significant.

Pairwise comparison	Alpha-diversity p-value (Shannon index; Wilcoxon)	Beta-diversity Bray-Curtis p-value (PERMANOVA)	Beta-diversity Jaccard p-value (PERMANOVA)
SARS-CoV-2 positive *vs* SARS-CoV-2 negative	0.37	0.42	0.642
SARS-CoV-2 positive (Ct value > 30) *vs* SARS-CoV-2 negative	0.35	0.398	0.621
NPS *vs* TS	**0.008**	**0.001**	**0.001**
Symptomatic *vs* asymptomatic	**<0.0001**	**0.001**	**0.001**
NPS-symptomatic-SARS-CoV-2 positive *vs* NPS-symptomatic SARS-CoV-2 negative	0.21	**0.036**	0.23
NPS-asymptomatic-SARS-CoV-2 positive *vs* NPS-asymptomatic-SARS-CoV-2 negative	0.21	0.162	**0.044**
TS-symptomatic-SARS-CoV-2	0.81	0.744	0.408

positive *vs* TS-symptomatic- SARS-CoV-2 negative	

Figure legends

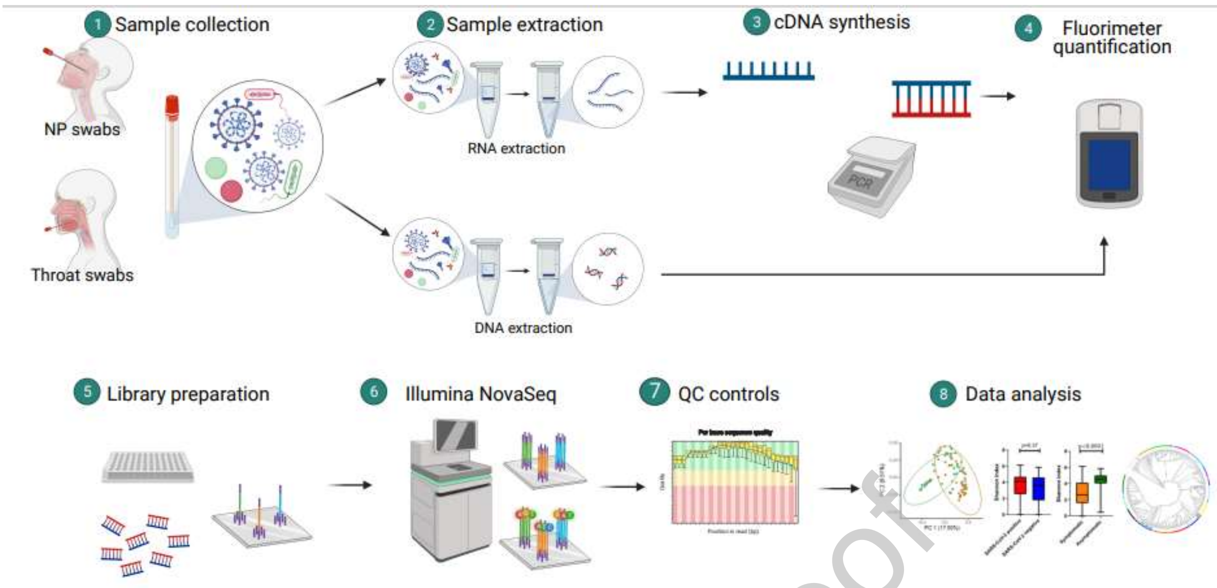


Figure 1. Metagenomics Next Generation Sequencing (mNGS) workflow. Nasopharyngal swabs (NPS) and throat swabs (TS) were collected from each patient. Both DNA and RNA were extracted independently from each sample. cDNA synthesis was performed from RNA extracts. Obtained purified and quantified dsDNA were submitted to library preparation followed by Illumina NovaSeq sequencing and subsequent data analysis.

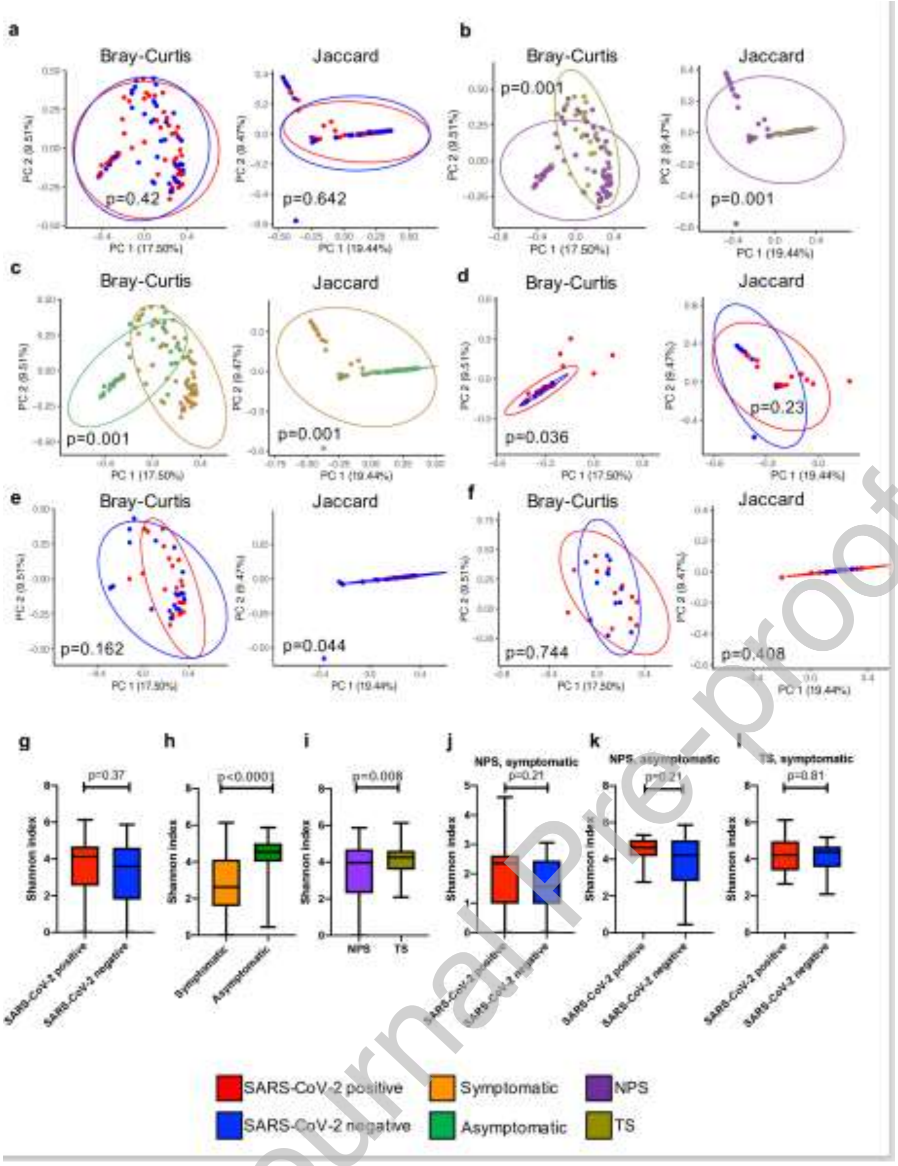


Figure 2. Alpha- and beta-diversity analysis of the DNA-metagenome. (a-f) Beta-diversity principal coordinate analysis (PCoA) plots were compared between the Bray-Curtis quantitative metric *vs* the Jaccard qualitative metric and a PERMANOVA was computed with 999 permutations in each pairwise comparison. (a) PCoAs by SARS-CoV-2 diagnostic status. (b) PCoAs by symptomatology status. (c) PCoAs by body site sampling. (d) PCoAs by SARS-CoV-2 diagnosis status amongst patients with

symptoms and NPS samples. (e) PCoAs by SARS-CoV-2 diagnosis status amongst asymptomatic patients and NPS samples. (f) PCoAs by SARS-CoV-2 diagnosis status amongst patients with symptoms and TS samples. (g-l) Alpha-diversity Shannon index with a pairwise Kruskal-Wallis test. (g) Shannon comparison by SARS-CoV-2 diagnostic status. (h) Shannon comparison by symptomatology status. (i) Shannon comparison by body site sampling. (j) Shannon comparison by SARS-CoV-2 diagnosis status amongst patients with symptoms and NPS samples. (k) Shannon comparison by SARS-CoV-2 diagnosis status amongst asymptomatic patients and NPS samples. (l) Shannon comparison by SARS-CoV-2 diagnosis status amongst patients with symptoms and TS samples. P-values below 0.05 were considered statistically significant.

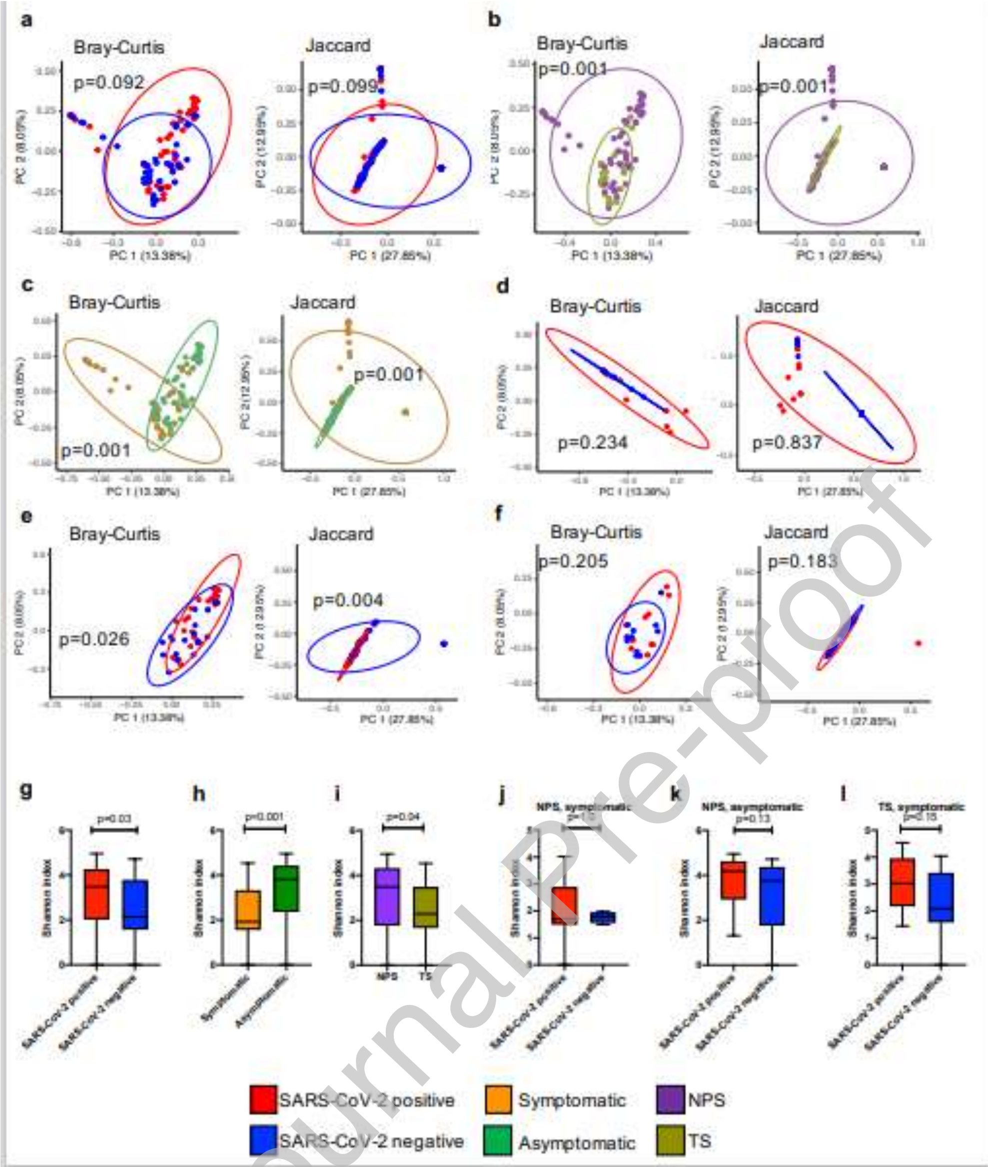


Figure 3. Alpha- and beta-diversity analysis of the bacteriophage RNA-metagenome. (a-f) Beta-diversity PCoAs were compared between the Bray-Curtis quantitative metric *vs* the Jaccard qualitative metric and a PERMANOVA was computed with 999 permutations in each pairwise comparison. (a) PCoAs by SARS-CoV-2 diagnostic status. (b) PCoAs by symptomatology status. (c) PCoAs by body site sampling. (d) PCoAs by SARS-CoV-2 diagnosis status amongst patients with symptoms and NPS

samples. (e) PCoAs by SARS-CoV-2 diagnosis status amongst asymptomatic patients and NPS samples. (f) PCoAs by SARS-CoV-2 diagnosis status amongst patients with symptoms and TS samples. (g-l) Alpha-diversity Shannon index with a pairwise Kruskal-Wallis test. (g) Shannon comparison by SARS-CoV-2 diagnostic status. (h) Shannon comparison by symptomatology status. (i) Shannon comparison by body site sampling. (j) Shannon comparison by SARS-CoV-2 diagnosis status amongst patients with symptoms and NPS samples. (k) Shannon comparison by SARS-CoV-2 diagnosis status amongst asymptomatic patients and NPS samples. (l) Shannon comparison by SARS-CoV-2 diagnosis status amongst patients with symptoms and TS samples.

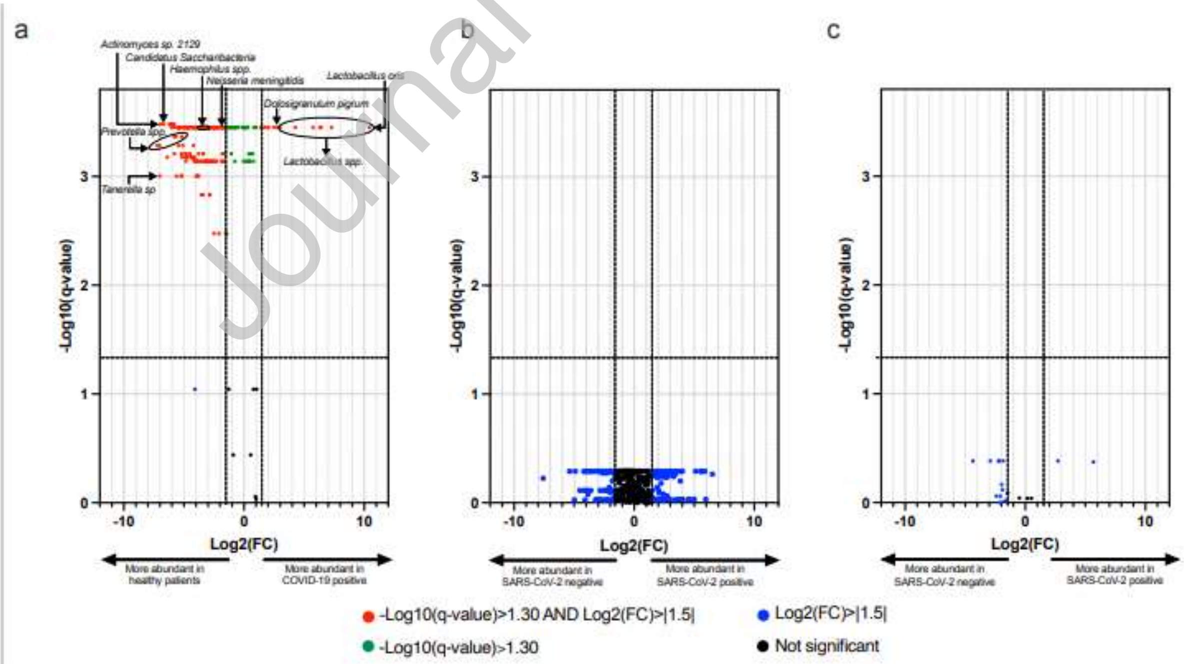


Figure 4. Volcano plots. Differences in species abundance between (a) individuals with COVID-19 from NPS samples and healthy patients from NPS samples, (b) NPS samples from SARS-CoV-2-infected asymptomatic patients and NPS samples from SARS-CoV-2-negative asymptomatic patients, and (c) NPS samples from SARS-CoV-2 infected symptomatic patients *vs* SARS-CoV-2-negative symptomatice individuals. P-values were obtained after performing a Wilcoxon-Mann-Whitney test and adjusted with the Benjamini-Hochberg correction. Vertical dashed lines represent a natural logarithm fold change of the mean of -1.5 and 1.5, respectively. The horizontal dashed line represents a -log10(q-value) of 1.30 (equivalent to a q-value of 0.05). Adjusted p-values below 0.05 were considered significant.

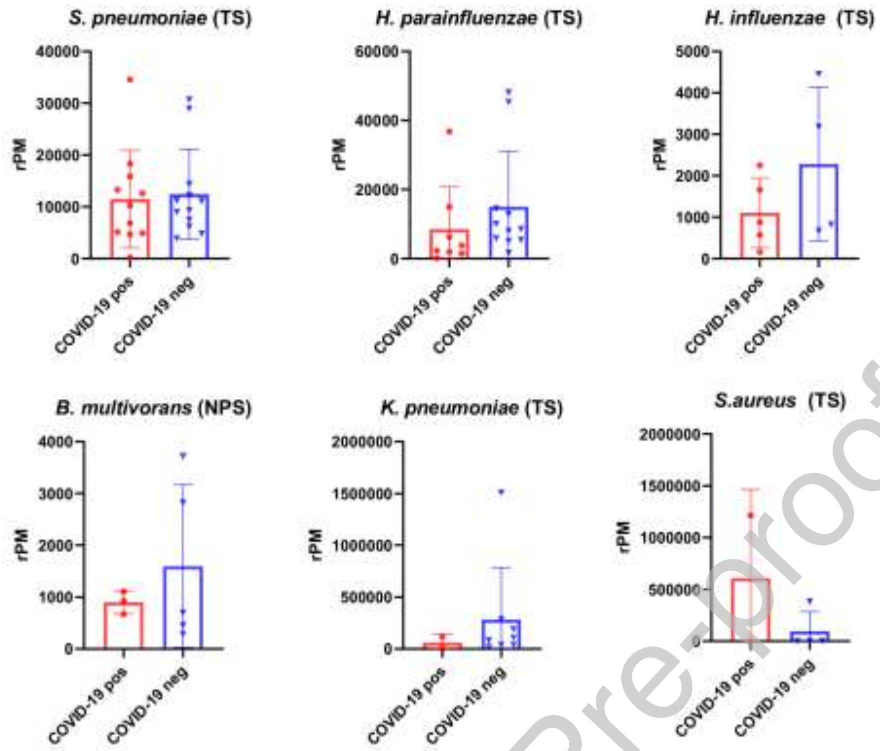


Figure 5. Comparison of rPM values of putative co-pathogens in symptomatic patients per COVID-19 status. Red dots=COVID positive patients; blue triangles=COVID negative patients. For each represented pathogen, the obtained rPM values after filtering are plotted for either throat swabs (TS) or nasopharyngal swabs (NPS). Mean values are plotted with corresponding standard deviation. Significance was assessed by performing a Wilcoxon-Mann-Whitney test. P-values below 0.05 were considered significant.

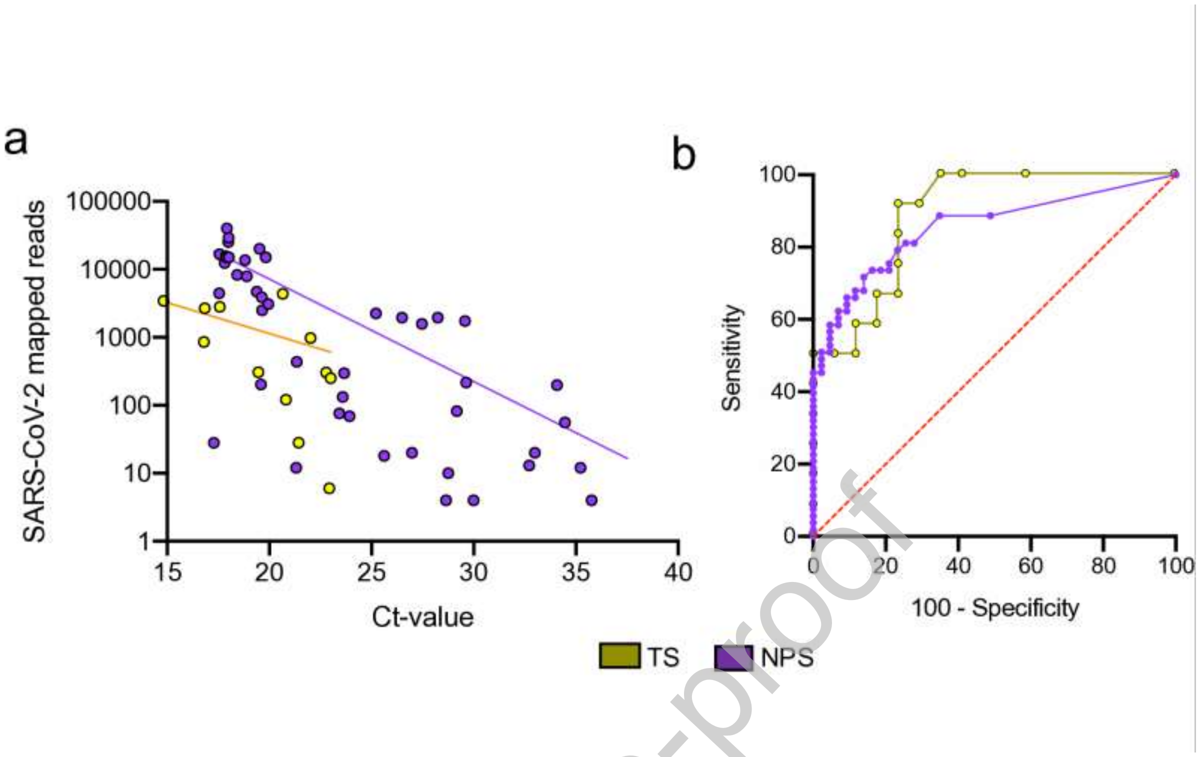


Figure 6. SARS-CoV-2 cDNA reads correlation with Egene real-time RT-PCR Ct values and ROC evaluation by anatomical sampling site. Number of SARS-CoV-2 cDNA reads are negatively correlated with RT-PCR Ct-value. (a) (Exponential regression of the mapped reads across all clinical isolates classified by anatomical swabbing site (n_{NPS}=48 and n_{TS}=12). Samples with no Ct-value were excluded from the analysis. (b) Mapped reads ROC curve of all clinical isolates classified by anatomical swabbing site (n_{NPS}=96 and n_{TS}=29).

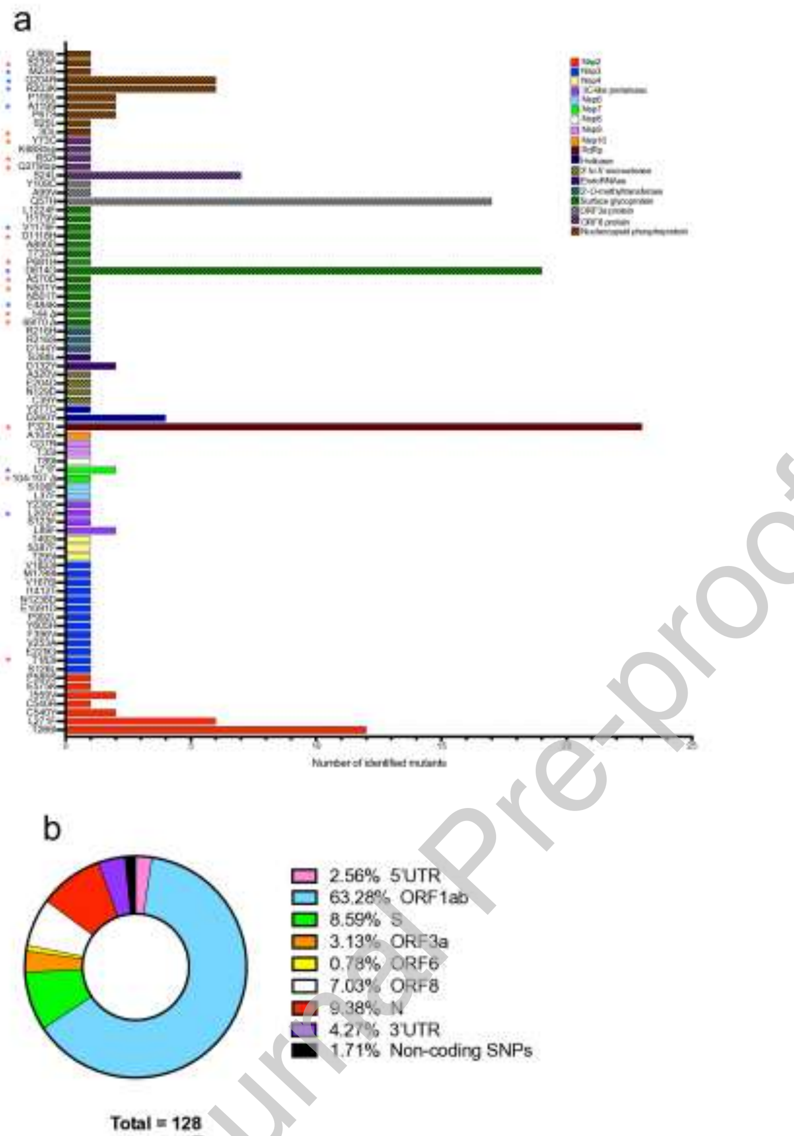


Figure 7. SARS-CoV-2 genomic variation. (a) Bar chart displaying the frequency of deletions and non-synonymous mutants. A blue asterisk depicts a non-synonymous mutant associated with the Zeta Brazilian VOI. A red asterisk depicts a non-synonymous mutant or deletion associated with the Alpha UK VOC. (b) Pie chart indicating the SARS-CoV-2 genomes identified SNPs (n=117) with a minimum of 10 nucleotides.

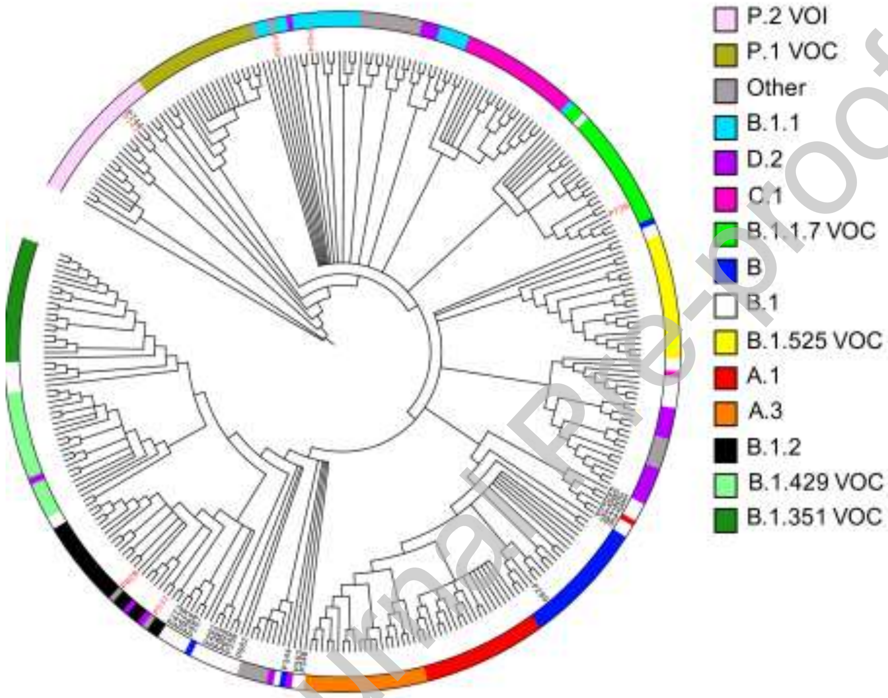


Figure 8. SARS-CoV-2 lineage identification

Maximum likelihood phylogenomics tree generated from 321 genomes taken from GISAID along with the consensus genome of each the clinical sample. Branches with bootstraps below 50 were collapsed to the next closest node. Clinical isolates are depicted with its patient ID. Black labels depict the samples which lineage was identified with Pangolin COVID-19 Lineage assigner. Red labels depict the samples in which lineage was inferred from the closest clade. The legend palette represents the corresponding SARS-CoV-2 lineage.